An On-Line Radio Access Technology Selection Algorithm in an LTE-WiFi Network

Arghyadip Roy, Prasanna Chaporkar, Abhay Karandikar
Department of Electrical Engineering, Indian Institute of Technology Bombay, India 400076
Email: {arghyadip, chaporkar, karandi}@ee.iitb.ac.in

***Abstract*—In a Heterogeneous Network (HetNet) comprising of multiple Radio Access Technologies (RATs), a user can be associated with a particular RAT and can be steered to other RATs in a seamless manner. To handle the rapid growth of data traffic, offloading of mobile data to Wireless Fidelity (WiFi) has been proposed in a Long Term Evolution (LTE) based HetNet. In this paper, we consider an optimal RAT selection problem in an offload-capable LTE-WiFi system with an objective of maximizing the total system throughput subject to a constraint on the voice user blocking probability. An on-line algorithm for optimal RAT selection is proposed based on a Relative Value Iteration Algorithm (RVIA) centric Q-learning approach. The proposed algorithm can be implemented without any explicit knowledge of arrival processes of voice and data users. Simulation results are presented to exhibit the convergence behavior of the proposed scheme to the optimal policy.**

I. Introduction

Various Radio Access Technologies (RATs) have been standardized [1] to meet the ever-increasing demands from users regarding different Quality of Services (QoSs). One of the serious challenges faced by cellular network operators today is a significant increase in data traffic. It has been predicted that the mobile data traffic will go beyond 15 exabytes per month by 2018 [2]. With such an alarming increase in mobile data traffic, interworking among different RATs has become essential from users' as well as network providers' perspective. In a Heterogeneous Network (HetNet) [1], where different RATs coexist, a user can be associated with a suitable RAT from a set of available RATs [1]. Our aim is to propose an optimal RAT selection scheme that does not require any system statistics like arrival rate of users.

In a conventional Third Generation Partnership Project (3GPP) Long Term Evolution (LTE)-based HetNet, LTE Base Stations (BSs) are deployed with an objective of providing ubiquitous coverage, while the IEEE 802.11 [3] Wireless Local Area Network (WLAN) (commonly known as Wireless Fidelity (WiFi)) Access Point (APs) with high bit rate capability are deployed mainly to provide coverage in hotspot areas. A user present in the dual coverage of an LTE BS and a WiFi AP can be associated with either of them. Additionally, after association, they can be steered to another RAT. This technique, known as mobile data offloading in an LTE-WiFi HetNet, has been introduced in 3GPP Release-12 [4] specification. Existing RAT selection and offloading algorithms can be typically classified into two categories, viz., user-initiated [1],[5]-[6] and network-initiated [7]-[10] algorithms. In [5], a user-initiated RAT selection algorithm is considered based on Signal-to-Noise Ratio (SNR) and load information of individual RATs inside an LTE-WiFi HetNet. Since all users make individual selfish decisions to maximize their utility functions, this algorithm may not result in a globally optimum solution. On the other hand, a network-initiated RAT selection algorithm optimizes different network parameters. In [7], a RAT selection problem is considered inside an LTE-WiFi HetNet with two user profiles. LTE is always preferred by high-priority users and only a portion of LTE resources can be shared with low-priority users. The objective is to maximize the operator revenue to obtain the optimal partitioning between the dedicated and shared resources. The network-initiated offloading approach described in [10] computes the optimal portion of traffic to offload to WiFi for maximizing the per-user system throughput.

System performance varies based on the chosen RAT selection policy. In this paper, our aim is to propose an optimal RAT selection policy for voice and data users in an LTE-WiFi HetNet. Offloading of data users is considered at the time of association of incoming users and departure of existing users, as the system state changes only at these points in time. We target to maximize the total throughput of the system. In our model, voice users are always associated with LTE as LTE can provide guaranteed QoS, whereas WiFi cannot satisfy the strict delay requirement associated with a voice user. On the other hand, although typically throughput offered by WiFi to data users is more than that of LTE, under high-traffic load, WiFi performance in terms of throughput may degrade. In such scenario, it may be preferable to associate data users with LTE instead of WiFi. Usually, data users in LTE provide more throughput to the overall system compared to the voice users. As a result, from the total system throughput perspective, association of a voice user with LTE may result in wastage of LTE resources, which may be better utilized by the data users. This may result in blocking of incoming voice users in LTE. To address the inherent trade-off between the throughput improvement and reduction of blocking probability of voice users, a constraint is considered on the voice user blocking probability. We formulate this as a constrained average reward Semi-Markov Decision Process (SMDP) problem, where the total system throughput is considered as the objective function subject to a voice user blocking probability constraint. Relative Value Iteration Algorithm (RVIA) is a well-known dynamic programming based technique to solve such problems but is

[1]The terminologies "association" and "RAT selection" are used interchangeably throughout this paper.

978-1-5090-4183-1/17/$31.00 ©2017 IEEE

computationally inefficient for large state spaces. In addition, the computation of optimal policy requires the transition probabilities between different states, which in turn require the knowledge of the arrival processes of data and voice users. It may be difficult to obtain in reality. Q-learning [11], which is a well-known Reinforcement Learning (RL) algorithm for MDP problems addresses this issue and can be implemented without the knowledge of the arrival processes. Very few of the existing works [12]-[13] focus on Q-learning algorithms for average reward MDPs and SMDPs, without requiring a separate timescale to update the average reward of the system.

The key contribution of this paper is to propose an RVIA based on-line Q-learning algorithm for a constrained average reward SMDP problem to compute the optimal RAT selection policy for an LTE-WiFi HetNet. The proposed algorithm does not require the knowledge of the arrival processes of voice and data users. We do not consider any separate timescale for the average reward. However, the Lagrange Multiplier (LM) associated with the constraint is updated along a different timescale. This work is an extension of our previous work [9], where a RAT selection problem in an LTE-WiFi HetNet has been formulated as an MDP problem with the parameters of the arrival processes known beforehand. Extensive simulations are conducted in ns-3 (a discrete event network simulator) [14] to evaluate the performance and convergence behavior of the proposed association algorithm. 3GPP recommended parameters are used throughout the simulations.

Although [6] and [8] undertake Q-learning based approaches for RAT selection and offloading inside an LTE-based HetNet, contrary to our network-initiated approach, these schemes are user-initiated in nature. The scheme in [8] makes use of network-provided information in taking RAT selection decision, without taking into account the possibility of offloading of data users. In [6], distributed traffic offloading decisions are taken by users based on the available local information, without any consideration of association strategies.

The rest of the paper is organized as follows. Section II describes the system model. In Section III, the problem is formulated within the framework of constrained SMDP. Section IV discusses the RL implementation for the problem described in Section III and proposes an on-line algorithm for RAT selection inside an LTE-WiFi HetNet. Section V presents simulation results. In Section VI, we conclude the paper.

II. System Model

We consider a system, where a WiFi AP with a relatively smaller coverage area is present within the coverage area of an LTE BS. Both the BS and the AP are assumed to be connected to a centralized controller by lossless links. We assume that voice and data users are present within the LTE coverage area. However, we consider only those data users which are present in the WiFi coverage area. There is a common resource pool in LTE for voice users and data users in the WiFi coverage area. Data users outside the coverage of WiFi always get associated with the LTE BS, and separate resources are assumed to be reserved for this purpose. All the users are assumed to be stationary.

A. State & Action Space

We assume that voice and data user arrivals follow Poisson processes with means λ_v and λ_d, respectively. Service times for voice and data user are exponentially distributed with means $\frac{1}{\mu_v}$ and $\frac{1}{\mu_d}$, respectively. The arrival of a new user to the system and departure of an existing user from the system are taken as decision epochs. Since the system is Markovian in nature, the system state is observed at each of the decision epochs. There is no need to consider the system state at other points in time as the system state does not change except at these epochs. An action chosen at every decision epoch depending on the current state moves the system to a different state with finite probability. Let the state in the state-space S be denoted by a vector $s = (x, y, z, w)$, where x, y denote the number of voice and data users in LTE and z denotes the number of data users in WiFi, respectively. The variable w takes the values $1, 2$ and 3 in the case of voice user arrival, data user arrival and voice user departure, respectively. In the case of data user departure from LTE and WiFi, w takes the values 4 and 5, respectively.

In a typical LTE network, voice and data users may be allocated different number of resource blocks from the resource pool. For the sake of simplicity, we assume that voice and data users have the same priority, and once admitted in LTE, are granted one unit of resource block each. It introduces the following constraints on the state-space S.

$$(x + y) \leq C; \quad z \leq L, \tag{1}$$

where C is the total number of common resource blocks reserved in LTE, and L is the maximum number of WiFi data users so that the per-user throughput in WiFi is more than a threshold (say 2 Mbps). Although the consideration of multiple resource blocks would replicate the real-life scenario more closely, this would complicate the underlying model without bringing any significant change in the methodology and approach adopted for this paper. Note that the above threshold is considered on the per-user throughput in WiFi, so that the average throughput experienced by users in WiFi does not drop too low.

Let the set of feasible actions in state s and the action space be denoted by $A(s)$ and A, respectively. Action 1 corresponds to blocking of an incoming user in case of an arrival or doing nothing at departure. Action 2 and 3 correspond to accepting an user in LTE and WiFi, respectively. Under action 4, a voice user is accepted in LTE and one data user is offloaded to WiFi. Action 5 offloads one data user to a RAT from which a departure has just occurred. When a voice user arrives into the system, an action from the set $\{1, 2, 4\}$ is selected. In the case of data user arrival, the feasible action set is $\{1, 2, 3\}$. Voice user blocking (action 1) is considered to be a feasible action in all states with $w = 1$. When a departure occurs, the controller chooses either action 1 or action 5. In this paper, we consider the blocking probability of voice users, defined as the fraction of blocked voice users, as a constraint for the considered problem. Since we consider the blocking of data users only when both LTE and WiFi reach their respective

capacities, we do not consider blocking probability of data users as a constraint.

From each state-action pair (s, a), the system moves to a different state s' with finite probability $p_{ss'}(a)$. In every state $s = (x, y, z, w)$, we denote the sum of arrival and service rates of all the users by $v(x, y, z)$. Thus, we have,

$$v(x, y, z) = \lambda_v + \lambda_d + x\mu_v + y\mu_d + z\mu_d. \tag{2}$$

For action a in state s,

$$p_{ss'}(a) = \begin{cases} \frac{\lambda_v}{v(x',y',z')}, & s' = (x', y', z', 1), \\ \frac{\lambda_d}{v(x',y',z')}, & s' = (x', y', z', 2), \\ \frac{x'\mu_v}{v(x',y',z')}, & s' = (x' - 1, y', z', 3), \\ \frac{y'\mu_d}{v(x',y',z')}, & s' = (x', y' - 1, z', 4), \\ \frac{z'\mu_d}{v(x',y',z')}, & s' = (x', y', z' - 1, 5). \end{cases} \tag{3}$$

Values of x', y' and z' as a function of arrivals and departures of users (the value of w) and action a are tabulated in Table I. An example of state transitions in different states

TABLE I: Transition Probability Table

(w, a)	(x', y', z')
$(\{1, \ldots 5\}, 1)$	(x, y, z)
$(1, 2)$	$(x + 1, y, z)$
$(2, 2)$	$(x, y + 1, z)$
$(2, 3)$	$(x, y, z + 1)$
$(1, 4)$	$(x + 1, y - 1, z + 1)$
$(\{3, 4\}, 5)$	$(x, y + 1, z - 1)$
$(5, 5)$	$(x, y - 1, z + 1)$

under different actions is demonstrated in Fig.1, where a data user arrival is followed by a voice user departure and a data user departure from WiFi, respectively. Actions taken corresponding to these states are illustrated in Fig.1.

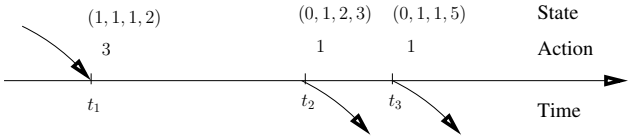


Fig. 1: An example of state transitions.

B. Rewards and Costs

Let the one-step reward and cost functions be represented by $r(s, a)$ and $c(s, a)$, respectively. Let $R_{L,V}$ and $R_{L,D}$ denote the bit rates of voice and data users in LTE, respectively. To keep our model simple and computationally tractable, we assume that $R_{L,D}$ is constant. Since, in practice, voice user generates traffic at Constant Bit Rate (CBR), we take $R_{L,V}$ (which is $\ll R_{L,D}$) to be a constant as well. $R_{W,D}(z)$ corresponds to the per-user data throughput of z users in WiFi, as considered in the full buffer WiFi model [15]. The calculation of $R_{W,D}(z)$ takes into account the contention-based medium access of WiFi users, i.e., the probabilistic transmission attempts of all the users in a slot, corresponding success and collision probabilities. $R_{W,D}(z)$ is also a function of slot times for successful transmission, idle slots and busy slots during collisions.

The reward function for a state-action pair is defined as the total throughput of the system in the state under the corresponding action. For example,

$$r(s, 2) = x.R_{L,V} + (y + 1).R_{L,D} + z.R_{W,D}(z). \tag{4}$$

Whenever the centralized controller in the system blocks an incoming voice user given that the LTE system is not saturated, the cost involved is one unit, else it is zero. Thus,

$$c(s, a) = \begin{cases} 1, & \text{if } s = \{(x, y, z, 1) \mid (x + y) \neq C\} \text{ and } a = 1, \\ 0, & \text{else.} \end{cases} \tag{5}$$

III. Problem Formulation

As discussed before, from the total system throughput perspective, since data users contribute more to the total throughput of the system, the association policy may result in blocking of incoming voice users. Our objective is to determine an association policy which maximizes the average expected total throughput of the system, subject to a constraint on the average expected blocking probability of voice users. Also, arrival and departure of users can occur at any point in time, making this a continuous time problem. Therefore, this problem can be formulated as a constrained SMDP.

A. Constrained SMDP Problem Formulation

Let $\mathbb{M}$ be the set of all memoryless policies. We assume that the Markov chains constructed under such policies are irreducible, to guarantee a unique stationary distribution. Let the infinite horizon average reward and cost of the system under the policy $M \in \mathbb{M}$ be denoted by V^M and B^M, respectively. Let $R(t)$ and $B(t)$ be the total reward and cost incurred up to time t, respectively.

For the constrained SMDP problem, our objective can be described as follows.

$$\begin{aligned} \text{Maximize:} \quad & V^M = \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}_M[R(t)], \\ \text{subject to:} \quad & B^M = \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}_M[B(t)] \leq B_{\max}, \end{aligned} \tag{6}$$

where $\mathbb{E}_M$ denotes the expectation operator when policy M is followed, and $B_{\max}$ denotes the constraint imposed upon the average expected blocking probability of voice users. Our objective is to determine the optimal policy for the constrained SMDP problem. Note that $\liminf$ and $\limsup$ in Equation (6) indicate the worst case reward and cost scenario.

B. Uniformization

To obtain the optimal policy, we need to apply RVIA [16]. However, first the SMDP need to be transformed into an equivalent discrete time MDP using uniformization [16]. After transformation into a discrete time model, the state space and action space remain the same. Let $\tau_s(a)$ represents the expected time until the next decision epoch, if the system is in state s and action a is chosen. The first step is to choose a number δ, such that $0 < \delta \leq \min_{s,a} \tau_s(a)$. Let $\hat{p}_{ss'}(a)$, $\hat{r}(s, a)$ and $\hat{c}(s, a)$ denote the transition probability, reward and cost

under action a in state s, respectively, in the transformed discrete time model. Thus, we have,

$$\hat{r}(s,a) = r(s,a)/\tau_s(a), \tag{7}$$

$$\hat{c}(s,a) = c(s,a)/\tau_s(a) \quad \text{and} \tag{8}$$

$$\hat{p}_{ss'}(a) = \begin{cases} \{\delta/\tau_s(a)\}p_{ss'}(a), & s' \neq s, \\ \{\delta/\tau_s(a)\}p_{ss'}(a) + [1 - \{\delta/\tau_s(a)\}], & s' = s. \end{cases} \tag{9}$$

C. Lagrangian Approach

After conversion into an equivalent discrete time MDP model, we use the Lagrangian approach [17] to solve. For a fixed value of LM β, the modified reward function is

$$\hat{r}(s,a;\beta) = \hat{r}(s,a) - \beta\hat{c}(s,a). \tag{10}$$

The dynamic programming equation described below provides the necessary condition for optimality for s, where $s' \in S$.

$$V(s) = \max_a [\hat{r}(s,a;\beta) + \sum_{s'} \hat{p}(s,s',a)V(s') - \rho], \tag{11}$$

where $V(s)$ denotes the value function of state $s \in S$ and ρ denotes the optimal average reward. For a fixed value of β, RVIA can be employed to solve the unconstrained maximization problem, as described below.

$$V_{n+1}(s) = \max_{a \in A(s)} [\hat{r}(s,a;\beta) + \sum_{s'} \hat{p}(s,s',a)V_n(s') - V_n(s^*)], \tag{12}$$

where $V_n(.)$, which converges to $V(.)$ for a large n, is an estimate of the value function after nth iteration and s^* is an arbitrary but fixed state. Next, the aim is to determine the value of β ($= \beta^*$, say) which maximizes the average expected reward, subject to the cost constraint. It is known that the considered problem has a stationary randomized optimal policy [17] for a particular value of $\beta = \beta^*$, with possible randomization in at most one state.

IV. Reinforcement Learning

In the previous section, the optimal policy can be obtained by employing RVIA, provided the transition probabilities $p_{ss'}(a)$ associated with a state-action pair are known beforehand. The knowledge of transition probability in turn requires the knowledge of the parameters of arrival processes of voice and data users. In practice, these parameters may be difficult to obtain. RL based techniques are good candidates in such scenarios. Typically, RL based techniques learn which action to perform by trial-and-error and hence can work in a model-free manner. We choose Q-learning [11] for its simplicity and popularity. However, since the model considered in this paper is a constrained SMDP model involving average reward, traditional Q-learning algorithm needs some modifications, as discussed below.

A. Proposed On-line Algorithm

The system moves through different states based on the arrival/departure of users and various actions taken in different states. Let the Q-value, the expected long-term average reward associated with a state s and action $\pi(s)$ as specified by the policy π, be denoted by $Q_\pi(s,\pi(s))$. The objective is to determine an optimal policy π^*, which maximizes the Q-value associated with a state, as described in the equation below.

$$\pi^*(s) = \arg\max_{a \in A(s)} Q_\pi(s,a), \quad \forall s, \pi. \tag{13}$$

Theory of stochastic approximation [18] enables us to remove the expectation in Equation (12) and achieve optimality in policy by doing averaging over time. Let $g(n)$ be an update sequence which possesses the following properties,

$$\sum_{n=1}^{\infty} g(n) = \infty; \sum_{n=1}^{\infty} (g(n))^2 < \infty. \tag{14}$$

Let $h(n)$ be another update sequence which possesses the same properties as described in Equation (14) along with the additional properties described below,

$$\sum_{n=1}^{\infty} (g(n) + h(n))^2 < \infty; \lim_{n\to\infty} \frac{h(n)}{g(n)} \to 0. \tag{15}$$

The key idea is to update the Q-value associated with one state-action pair at a time and keep the other values unchanged. This scheme translates into the following equation.

$$\begin{aligned} &Q_{n+1}(s,a) = (1 - g(n))Q_n(s,a) + g(n)[r(s,a) - \\ &\beta c(s,a) + \max_{a' \in A(s')} Q_n(s',a') - Q_n(s^*,a^*)t(s,a,s')]; \\ &Q_{n+1}(\tilde{s},\tilde{a}) = Q_n(\tilde{s},\tilde{a}) \quad \forall (\tilde{s},\tilde{a}) \neq (s,a), \end{aligned} \tag{16}$$

where $t(s,a,s')$ denotes the random transition time to move from state s to s' under the action a and (s^*,a^*) is a fixed state-action pair. However, this scheme works for a fixed value of LM β. To obtain the optimal value of β, β is to be iterated along the timescale $h(n)$, as specified below.

$$\beta_{n+1} = \Lambda[\beta_n + h(n)(B_n - B_{\max})], \tag{17}$$

where the projection operator Λ ensures that the value of LM remains bounded in the interval $[0, L]$ for a large $L > 0$. The assumptions on $g(n)$ and $h(n)$ as specified in Equation (14) and (15) guarantee that two variables are updated on two different timescales. In addition, the update of LM is done on a slower timescale than the update of Q-value of state-action pair. From the slower LM timescale perspective, $Q(s,a)$ appears to be converged to optimality corresponding to the current LM value, and from a faster timescale perspective, LM appears to be almost fixed. $g(n)$ and $h(n)$ are two different learning rates, which specify how much importance is to be given to the old Q-value over the present received reward.

At every decision epoch, i.e., upon every arrival or departure of users, the centralized controller chooses an action. If the network receives a high reward by selecting an action, it may prefer to *exploit* that action in future decision epochs. However, the network needs to *explore* other actions as well

with time to observe whether they result in a significant amount of rewards. The aim of the proposed algorithm is to exploit the actions with high reward with a sufficient number of explorations. In this paper, we adopt the ϵ-greedy [11] approach for exploration and exploitation. At every decision epoch, if the network is in state s, network explores with probability $\epsilon(s)$ and exploits the action having the highest Q-value with probability $1-\epsilon(s)$. In exploration phase, all the feasible actions in state s are chosen with equal probabilities, and exploration is gradually reduced over time.

The two timescale Q-learning algorithm proposed is described in Algorithm 1. As described in the algorithm, Q-values associated with different state-action pairs, the LM and the number of iterations are initialized at the beginning. Based on a random event (arrival or departure), the system state is initialized. When the system is in state s, it chooses exploration and exploitation with finite probabilities. In exploration phase, a random action is selected, while in exploitation phase, the system chooses the action with the highest Q-value. Based on the observed reward in that state and the transition time to next state s', $Q(s,a)$ is updated along with the LM. This process is thus continued for all decision epochs. The simulation results demonstrate that this algorithm indeed converges in a reasonable number of iterations, provided the number of visits to each state-action pair is sufficiently large, and the learning rates are slowly reduced to zero, as specified in Equation (14).

Algorithm 1 Constrained SMDP based two timescale Q-learning algorithm

1: Initialize number of iterations $k \leftarrow 1$, Q-values of state-action pairs $Q(s,a) \leftarrow 0, \quad \forall s \in S, a \in A(s)$ and the LM $\beta \leftarrow 0$.
2: **while** TRUE **do**
3: Determine the system state s.
4: **if** exploration phase **then**
5: Choose one of the feasible actions at random.
6: **else**
7: Choose action $a = \arg\max_a Q(s,a)$.
8: **end if**
9: Observe reward $r(s,a;\beta) = r(s,a) - \beta c(s,a)$.
10: Go to next state s'.
11: Observe transition time to next state $t(s,a,s')$.
12: Update $Q(s,a)$ according to Equation (16).
13: Update the LM according to Equation (17).
14: Update $s \leftarrow s'$ and $k \leftarrow k+1$.
15: **end while**

V. Simulation Results

In this section, the algorithm proposed in the last section is simulated in ns-3 to observe the performance. Optimal policy obtained by solving the SMDP problem using RVIA, when the arrival rates of voice and data users are known, is also simulated in ns-3. It is observed that Algorithm 1 converges to the optimal policy as the time progresses. Moreover, the performance of Algorithm 1 in terms of total system throughput is compared to the performance of the optimal policy.

A. Simulation Model and Evaluation Procedure

The network model is simulated with a 3GPP LTE BS and an IEEE 802.11g WiFi AP. All users are assumed to be stationary. The AP is trusted from the point of view of interworking and approximately 50 m away from the LTE BS. Data users are assumed to be distributed uniformly within 30 m radius of the WiFi AP. LTE and WiFi network parameters used in our simulation are based on 3GPP [19]-[20] models and saturation throughput [15] of IEEE 802.11g WiFi [3] model, respectively, as summarized in Table II and III. Although the system model does not have any assumption on the scheduling strategy, for simulation purposes, we consider proportional fair scheduling for the LTE BS. For the Q-value and LM updates, we consider $g(n) = \frac{1}{n^{0.5}}$ and $h(n) = \frac{1}{n}$.

TABLE II: LTE Network Model

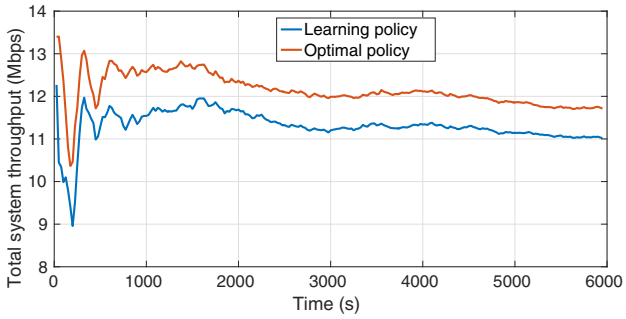
Parameter	Value
Maximum capacity	10 users
Voice bit rate of a single user	20 kbps
Data bit rate of a single user	5 Mbps
Voice packet payload	50 bits
Data packet payload	600 bits
Tx power for BS and MS	46 dBm and 23 dBm
Noise figure for BS and MS	5 dB and 9 dB
Antenna height for BS and MS	32 m and 1.5 m
Antenna parameter for BS and MS	Isotropic Antenna
Path loss	$128.1 + 37.6\log(R)$, R in kms

TABLE III: WiFi Network Model

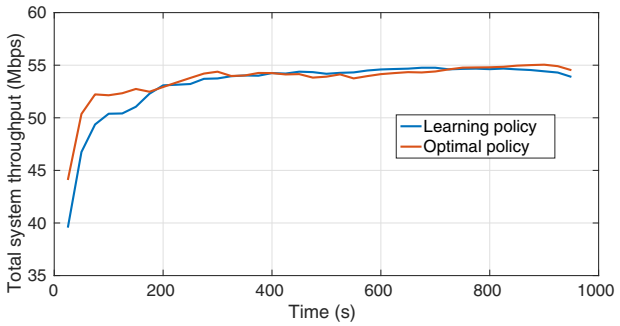
Parameter	Value
Channel bit rate	54 Mbps
UDP header	224 bits
Packet payload	1500 bytes
Slot duration	20 μs
Short inter-frame space (SIFS)	10μs
Distributed Coordination Function IFS (DIFS)	50μs
Minimum acceptable per-user throughput	3.5 Mbps
Tx power for AP	23dBm
Noise figure for AP	4 dB
Antenna height for AP	2.5 m
Antenna parameter	Isotropic antenna

B. Convergence Analysis

Fig.(2a) and (2b) illustrate how Algorithm 1 converges to the optimal policy in terms of the total system throughput. It is evident that as the time progresses, the total throughput of the system gradually becomes closer to the total throughput under the optimal policy. However, in case of Fig.(2a), since the users are serviced at faster rates than the corresponding arrival rates ($\lambda_v < \mu_v$ and $\lambda_d < \mu_d$) and the simulations start with an initially empty system, the average number of voice and data users in the system at any point in time is less. As a result, all the states in the state space are not visited often. Even after a sufficient number of decision epochs, occasionally some new state is visited, and the algorithm starts to learn the optimal policy in that state by trial-and-error. Therefore, even after a sufficient amount of time, the total throughput provided by Algorithm 1 is marginally lower (around 0.6 Mbps) than that of the optimal policy. Nevertheless, the performance offered by the learning algorithm is very close to optimal in the long run. On the contrary, in case of $\lambda_v = \lambda_d = 1.0s^{-1}$ and $\mu_v =$



(a) $\lambda_v = \lambda_d = 0.5s^{-1}$ and $\mu_v = \mu_d = 1s^{-1}$.



(b) $\lambda_v = \lambda_d = 1.0s^{-1}$ and $\mu_v = \mu_d = 0.1s^{-1}$.

Fig. 2: Comparison of performance of optimal policy and learning based policy.

$\mu_d = 0.1s^{-1}$ (See Fig.(2b)), since the arrival rates are higher than the service rates, the probability of visiting all the states in the state space is relatively higher. Hence, after a sufficient amount of time, almost all the state-action pairs are visited a reasonable number of times, resulting in convergence to the optimal policy. Fig.(2b) demonstrates that indeed optimality in total network throughput is achieved after almost $200s$. Except this small interval at the beginning of the simulation, in most of the states, the action with the highest Q-value matches with the action chosen by the optimal policy.

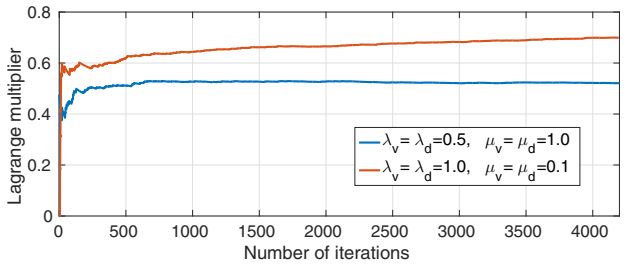


Fig. 3: Convergence of Lagrange multiplier.

In Fig.3, we demonstrate how the value of LM converges with the number of iterations under different values of $\lambda_v, \lambda_d, \mu_v$ and μ_d. Upon every arrival and departure, the value of LM is updated as specified in Algorithm 1, depending on the cost incurred by the corresponding action chosen. It is observed that for the two cases we consider, the LM converges approximately after 500 and 1500 iterations, respectively. However, the actual time taken to achieve convergence depends on the chosen value of arrival and service rates.

VI. Conclusions

In this paper, an on-line learning algorithm has been proposed for obtaining the optimal RAT selection policy in an offload-capable LTE-WiFi HetNet. This algorithm does not require the knowledge of arrival processes of data and voice users and hence fits well for an on-line implementation under the paradigm of stochastic approximation. A modified Q-learning based approach has been adopted for the proposed algorithm. Simulation results have demonstrated that this algorithm indeed converges to optimality after a reasonable number of iterations.

VII. Acknowledgment

This work is funded by the Department of Electronics and Information Technology (DeitY), Government of India.

References

[1] F. Rebecchi, M. D. de Amorim, V. Conan, A. Passarella, R. Bruno and M. Conti, "Data Offloading Techniques in Cellular Networks: a Survey," IEEE Communications Surveys & Tutorials, Vol. 17, No. 2, pp. 580-603, Nov 2014.

[2] Cisco, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, pp. 2013–2018, Feb 2014.

[3] IEEE 802.11-2012, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Mar 2012.

[4] 3GPP TR 37.834 V0.3.0, "Study on WLAN/3GPP Radio Interworking," May 2013.

[5] M. Gerasimenko, N. Himayat, S.P. Yeh, S. Talwar, S. Andreev and Y. Koucheryavy, "Characterizing Performance of Load-aware Network Selection in Multi-radio (WiFi/LTE) Heterogeneous Networks," in GLOBECOM Workshop, pp. 397-402, Dec 2013.

[6] K. Adachi, M. Li, P. H. Tan, Y. Zhou and S. Sun, "Q-Learning Based Intelligent Traffic Steering in Heterogeneous Network," in proc. of VTC Spring, pp. 1-5, May 2016.

[7] E. Khloussy, X. Gelabert and Y. Jiang, "A Revenue-Maximizing Scheme for Radio Access Technology Selection in Heterogeneous Wireless Networks with User Profile Differentiation," Advances in Communication Networking, Springer, pp. 66-77, 2013.

[8] M. El Helou, M. Ibrahim, S. Lahoud, K. Khawam, D. Mezher and B. Cousin, "A Network-assisted Approach for RAT Selection in Heterogeneous Cellular Networks," IEEE Journal on Selected Areas in Communications, Vol. 33, No. 6, pp. 1055–1067, Jun 2015.

[9] A. Roy and A. Karandikar "Optimal Radio Access Technology Selection Policy for LTE-WiFi Network," in proc. of WiOpt, pp. 291 - 298, May 2015.

[10] B. H. Jung, N. Song and D. K. Sung, "A Network-assisted User-centric WiFi-Offloading Model for Maximizing Per-user Throughput in a Heterogeneous Network," IEEE Transactions on Vehicular Technology, Vol. 63, Issue 99, pp. 1940 - 1945, Oct 2013.

[11] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, Cambridge: MIT Press, 1998.

[12] J. Abounadi, D. Bertsekas and V. S. Borkar, "Learning Algorithms for Markov Decision Processes with Average Cost," SIAM Journal on Control and Optimization, Vol. 40, No. 3, pp. 681-698, 2001.

[13] A. Gosavi, "Relative Value Iteration for Average Reward Semi-Markov Control via Simulation," in proc. of WSC, pp. 623-630, Dec 2013.

[14] http://code.nsnam.org/ns-3-dev/.

[15] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," IEEE Journal on Selected Areas in Communications, Vol. 18, No. 3, pp. 535-547, Mar 2000.

[16] M. L. Putterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley and Sons, 1994.

[17] E. Altman, Constrained Markov Decision Processes, CRC Press, 1999.

[18] V. S. Borkar, Stochastic Approximation, Cambridge Books, 2008.

[19] 3GPP TR 36.814 V9.0.0, "Further Advancements for E-UTRA Physical Layer Aspects," Mar 2010.

[20] 3GPP TR 36.839 V11.1.0, "Mobility Enhancements in Heterogeneous Networks," Dec 2012.